

ON THE NEIGHBOURHOODS OF TREES

PETER J. HUMPHRIES AND TAOYANG WU

ABSTRACT. Tree rearrangement operations typically induce a metric on the space of phylogenetic trees. One important property of these metrics is the size of the neighbourhood, that is, the number of trees exactly one operation from a given tree. We present an expression for the size of the TBR (tree bisection and reconnection) neighbourhood, thus answering a question first posed in [1].

1. INTRODUCTION

Phylogenetic trees are a commonly used tool for representing the relationships between species in an evolutionary system, especially in evolutionary biology. A central task in the study of these trees is to determine which among a set of hypothesised trees gives the best explanation of empirical data. However, finding the trees that optimize some criterion is often computationally prohibitive because of the large number of trees to be checked. An approach that avoids this is a heuristic hill-climbing algorithm that searches tree space using tree rearrangement operations [5, 6]. That is, at each iteration the optimal tree within one rearrangement operation is chosen as the input for the next step, and the algorithm is thus guaranteed to find a local optimum.

Loosely speaking, a tree rearrangement operation breaks a tree into two contiguous parts, and rejoins these parts to form a new tree. Among the three tree rearrangement operations of interest, namely NNI (nearest neighbour interchange), SPR (subtree prune and regraft) and TBR (tree bisection and reconnection), each induces a distinct metric on the space of unrooted trees. Several properties of these metrics are

Date: February 13, 2012.

Key words and phrases. tree rearrangement, TBR, phylogenetic tree.

The authors thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, where this work began. Peter Humphries was supported by the New Zealand Marsden Fund. Taoyang Wu thanks Queen Mary, University of London, where he undertook some of this research. He was also supported in part by the Engineering and Physical Sciences Research Council (EP/D068800/1). A preliminary version of this paper is included in the authors' respective Ph.D. dissertations. Finally, we thank Dr. Arnau Mir Torres and two anonymous referees for their valuable comments.

important for understanding the efficiency of the algorithm outlined above. Our interest in this paper is in the size of the TBR neighbourhood, that is, the number of trees that can be reached from a specified starting tree via a single TBR operation.

A phylogenetic tree is an unrooted binary tree in the graph theoretic sense, with a unique label attached to every leaf, or vertex of degree one. We denote by \mathcal{T}_n the collection of all phylogenetic trees whose leaves are the set $\{1, \dots, n\}$.

For a tree $\mathcal{T} \in \mathcal{T}_n$, where $n \geq 4$, Robinson [7] showed that the NNI neighbourhood $N_{\text{NNI}}(\mathcal{T})$ has size exactly equal to $2n - 6$, that is,

$$|N_{\text{NNI}}(\mathcal{T})| = 2n - 6,$$

while Allen and Steel [1] proved that

$$|N_{\text{SPR}}(\mathcal{T})| = 2(n - 3)(2n - 7),$$

where $N_{\text{SPR}}(\mathcal{T})$ is the SPR neighbourhood of \mathcal{T} . It was also demonstrated in [1] that the size of the TBR neighbourhood is dependent on the shape of \mathcal{T} . More recently, in [3], the bounds

$$cn^2 \log n + O(n^2) \leq |N_{\text{TBR}}(\mathcal{T})| \leq \frac{2}{3}n^3 - 4n^2 + \frac{16}{3}n + 2$$

were shown to hold for all $n \geq 4$, with the upper bound being met with equality if and only if \mathcal{T} is a caterpillar, that is, a phylogenetic tree in which every non-leaf vertex is adjacent to a leaf.

The rest of this paper is divided into three sections. Section 2 contains the definitions required to follow the main content of the paper. In Section 3, we relate the number of possible rearrangement operations for \mathcal{T} to the size of the neighbourhood, and use this to reprove Allen and Steel's [1] result for the SPR neighbourhood and to obtain an expression for the TBR neighbourhood dependent on the tree shape. In Section 4, we characterise the trees that respectively maximise and minimise the size of $N_{\text{TBR}}(\mathcal{T})$ for all binary tree spaces \mathcal{T}_n . These characterisations are also extended to reprove the tight upper bound given in [3], and to further prove an asymptotically tight lower bound.

2. DEFINITIONS

Before giving formal definitions of each rearrangement operation, we introduce some useful terminology. Given a tree \mathcal{T} and a subset X of the leaf set of \mathcal{T} , the restriction of \mathcal{T} to X , or $\mathcal{T}|X$, is the minimal subtree of \mathcal{T} connecting the leaves in X , with all vertices of degree two suppressed. A split $X|Y$ of a tree is a bipartition of the leaf set such that $\mathcal{T}|X$ and $\mathcal{T}|Y$ are vertex disjoint subtrees of \mathcal{T} . Further, if

$X' \subseteq X, Y' \subseteq Y$, then we call $X'|Y'$ a partial split of \mathcal{T} . A split is trivial if one of its parts contains only one leaf. The set of all splits of \mathcal{T} is denoted by $\Sigma(\mathcal{T})$. If \mathcal{T} is a tree with the leaf set Z , then a cluster of \mathcal{T} is a set X such that $X|(Z - X) \in \Sigma(\mathcal{T})$. If $|X| = 2$, then we call X a cherry.

A binary tree is a tree whose vertex degree is either one or three. Note that a binary tree with $n \geq 3$ leaves has $2n - 2$ vertices in total and $2n - 3$ edges, an observation that will be used throughout this paper.

Although NNI was the point of departure for the study of these operations [7], we will first define TBR, being the most general of the three. A TBR operation on a binary phylogenetic tree \mathcal{T} involves deleting some edge e from \mathcal{T} (bisection), and subsequently inserting a new edge f so that the resulting tree \mathcal{T}' is distinct from \mathcal{T} (reconnection). Since we require \mathcal{T}' to be binary, it is necessary to subdivide an edge in one (in the case that the other component is an isolated labelled vertex) or both components created in the bisection stage before inserting the new edge. An example is given in Fig. 1. We can transform \mathcal{T}_1 into \mathcal{T}_2 by first deleting the edge e from \mathcal{T}_1 , and then adding the new edge f . To check that there has been no other change to the tree's structure, note that deleting e from \mathcal{T}_1 gives the same forest as deleting f from \mathcal{T}_2 .

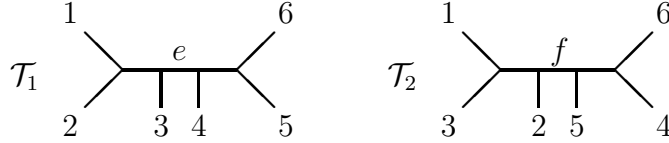


FIGURE 1. Two trees $\mathcal{T}_1, \mathcal{T}_2 \in \mathcal{T}_6$ that are one TBR operation apart.

For a binary tree \mathcal{T} , we define the set $\mathcal{O}_{\text{TBR}}(\mathcal{T})$ to be all possible TBR operations θ that can be applied to the tree \mathcal{T} . An important point to note here is that for distinct $\theta_1, \theta_2 \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$, we may have $\theta_1(\mathcal{T}) = \theta_2(\mathcal{T})$. The reason for this is that an operation $\theta \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ is not specified solely by the output tree $\theta(\mathcal{T})$, but also by the edge e that is deleted from \mathcal{T} in the bisection stage of θ .

Observe that for any two distinct trees $\mathcal{T}, \mathcal{T}' \in \mathcal{T}_X$, there is a TBR operation $\theta \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ for which $\theta(\mathcal{T}) = \mathcal{T}'$ if and only if there is some split $X_1|X_2 \in \Sigma(\mathcal{T}) \cap \Sigma(\mathcal{T}')$ such that $\mathcal{T}|X_i = \mathcal{T}'|X_i$ for all $i \in \{1, 2\}$. In this case, $X_1|X_2$ is the split induced by θ . To demonstrate this, if the

edges e and f have respectively been deleted and inserted in the TBR operation that changes \mathcal{T} into \mathcal{T}' , then the forest obtained by deleting e from \mathcal{T} must be identical to the forest obtained by deleting f from \mathcal{T}' . This provides not only the common bipartition of the leaf set, but also the common subtrees induced by each part of this bipartition.

SPR is a special case of TBR in which there is less freedom at the reconnection stage. Let \mathcal{T} be a binary tree, and let $\theta \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ be a TBR operation on \mathcal{T} in which the edge e is deleted, and let $X_1|X_2$ be the split of \mathcal{T} induced by e . Then θ is an SPR operation for \mathcal{T} if and only if, without loss of generality, $\mathcal{T}|(X_2 \cup x_1) = \theta(\mathcal{T})|(X_2 \cup x_1)$ for some $x_1 \in X_1$. Moreover, if this holds then in fact the same property holds for all $x_1 \in X_1$.

The significance of this condition is that one of the components formed in the bisection of \mathcal{T} , in this case $\mathcal{T}|X_2$, is treated as a rooted subtree, and is then regrafted so that this rooting is preserved with respect to the other component. We say that we have pruned $\mathcal{T}|X_2$ from \mathcal{T} , and regrafted it to form \mathcal{T}' .

The previous example (refer to Fig. 1) does not represent an SPR operation, since neither component obtained by deleting e from \mathcal{T}_1 can be regrafted to the other to form \mathcal{T}_2 . By making a subtle change, in particular by exchanging the labels 4 and 5 on \mathcal{T}_2 , we get a tree \mathcal{T}_3 that can be obtained from \mathcal{T}_1 by a single SPR operation. This example is shown in Fig. 2.

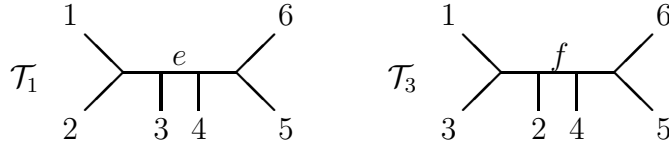


FIGURE 2. Two trees $\mathcal{T}_1, \mathcal{T}_3 \in \mathcal{T}_6$ that are one SPR operation apart.

NNI operations are TBR operations in which the reconnection is still more restricted than for SPR. Let \mathcal{T} be a phylogenetic tree, and let $\theta \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ be an SPR operation in which $\mathcal{T}|Y$ is pruned from \mathcal{T} and regrafted to form $\mathcal{T}' = \theta(\mathcal{T})$. We say that θ is an NNI operation if and only if there is some cluster $Z \neq Y$ of \mathcal{T} such that we can form \mathcal{T}' from \mathcal{T} by swapping the subtrees $\mathcal{T}|Y$ and $\mathcal{T}|Z$. In this case, $\mathcal{T}|Y$ and $\mathcal{T}|Z$ can be seen as adjacent in some sense, as shown by the schematic diagram in Fig. 3. Note that \mathcal{T}' can be obtained from \mathcal{T} by four distinct NNI operations, namely pruning one of the subtrees in

$\{\mathcal{T}|X, \mathcal{T}|Y, \mathcal{T}|Z, \mathcal{T}|W\}$ and regrafting it in an appropriate way. Indeed, if θ is an NNI operation for \mathcal{T} , then there are precisely four distinct operations $\theta' \in \mathcal{O}_{\text{NNI}}(\mathcal{T})$ such that $\theta(\mathcal{T}) = \theta'(\mathcal{T})$. The possibility that two distinct operations can result in the same tree lies behind the main lemma (Lemma 3.1) in Section 3.

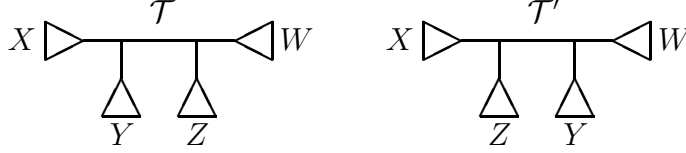


FIGURE 3. Two trees that are one NNI operation apart.

Extending our earlier notation for TBR to both SPR and NNI, we have

$$\mathcal{O}_{\text{NNI}}(\mathcal{T}) \subseteq \mathcal{O}_{\text{SPR}}(\mathcal{T}) \subseteq \mathcal{O}_{\text{TBR}}(\mathcal{T})$$

for any tree \mathcal{T} . The TBR neighbourhood of \mathcal{T} is the set

$$N_{\text{TBR}}(\mathcal{T}) = \{\theta(\mathcal{T}) : \theta \in \mathcal{O}_{\text{TBR}}(\mathcal{T})\}.$$

That is, $N_{\text{TBR}}(\mathcal{T})$ is the set of all trees that are precisely one TBR rearrangement operation from \mathcal{T} . The NNI neighbourhood $N_{\text{NNI}}(\mathcal{T})$ and the SPR neighbourhood $N_{\text{SPR}}(\mathcal{T})$ are defined similarly. Clearly, the elements in these neighbourhoods are dependent on the operation in question, and we have the corresponding nesting property as above. More explicitly,

$$N_{\text{NNI}}(\mathcal{T}) \subseteq N_{\text{SPR}}(\mathcal{T}) \subseteq N_{\text{TBR}}(\mathcal{T}).$$

3. NEIGHBOURHOOD SIZES

The approach used by Allen and Steel [1] to determine both the size of the SPR neighbourhood and the upper bound on the size of the TBR neighbourhood was to count directly the number of trees that can be obtained from \mathcal{T} via a single operation. While this seems the most natural approach, there is a fundamental barrier to performing this enumeration that we alluded to briefly in Section 2. This is the fact that some operations in $\mathcal{O}_{\text{TBR}}(\mathcal{T})$ may be redundant. That is, there may be distinct elements $\theta_1, \theta_2 \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ for which

$$\theta_1(\mathcal{T}) = \theta_2(\mathcal{T}).$$

This potentially leads to counting some trees in $N_{\text{TBR}}(\mathcal{T})$ more than once. If we can determine precisely which operations in $\mathcal{O}_{\text{TBR}}(\mathcal{T})$ output

the same tree, then we can relate the size of the TBR neighbourhood to the number of operations on \mathcal{T} .

It transpires, as the next lemma shows, that the only redundant TBR operations are all NNI operations.

Lemma 3.1. *Let $\theta, \theta' \in \mathcal{O}_{\text{TBR}}(\mathcal{T})$ be distinct TBR operations. If $\theta(\mathcal{T}) = \theta'(\mathcal{T})$, then $\theta \in \mathcal{O}_{\text{NNI}}(\mathcal{T})$.*

Proof. Suppose that $A|B$ is the split of \mathcal{T} induced by θ , and that $A'|B'$ is the split induced by θ' . Then $A|B \neq A'|B'$ as otherwise $\theta(\mathcal{T})$ must be distinct from $\theta'(\mathcal{T})$. Hence we may assume that $A \subset A'$, and hence also $B' \subset B$. Since $\mathcal{T}|A' = \theta(\mathcal{T})|A'$, we have immediately that $\theta \in \mathcal{O}_{\text{SPR}}(\mathcal{T})$. Let $A_0 = A, A_1, \dots, A_k = A'$ be clusters of \mathcal{T} such that

- (i) $A_i|B'$ is a partial split of \mathcal{T} ; and
- (ii) A_{i+1} is a minimal cluster of \mathcal{T} that contains A_i .

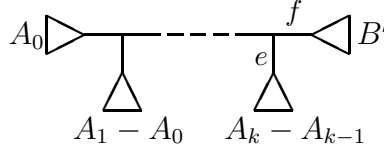


FIGURE 4. The tree \mathcal{T} in the proof of Lemma 3.1.

The generic structure of \mathcal{T} is depicted in Fig. 4, where f and e are two edges and whose removal will result in the split $A'|B'$ and $(A_k - A_{k-1})|(X - (A_k - A_{k-1}))$, respectively. Now consider the operation θ . If $k \geq 3$ then in order for $\mathcal{T}|A' = \theta(\mathcal{T})|A'$ to hold, we must regraft the pruned subtree $\mathcal{T}|A$ in the same place, but this implies $\mathcal{T} = \theta(\mathcal{T})$, a contradiction. If $k = 2$, to ensure $\mathcal{T}|A' = \theta(\mathcal{T})|A'$, we must regraft $\mathcal{T}|A$ to the edge e or f . In other words, $\theta(\mathcal{T})$ is obtained from \mathcal{T} by swapping either the subtrees $\mathcal{T}|A$ and $\mathcal{T}|B'$, or $\mathcal{T}|A$ and $\mathcal{T}|(A_2 - A_1)$, from which it follows that θ is an NNI operation. Now it remains to establish the case $k = 1$. To this end, we can further assume that $|A_1 - A_0| > 1$, because otherwise $\mathcal{T}|A' = \theta(\mathcal{T})|A'$ implies the contradiction $\mathcal{T} = \theta(\mathcal{T})$. Therefore the generic structure of \mathcal{T} in this case can be represented as in Fig. 5, where $C_1 \cup C_2 = A_1 - A_0$. Using the constraint $\mathcal{T} \neq \theta(\mathcal{T})$ and $\mathcal{T}|A' = \theta(\mathcal{T})|A'$ again, we can assert that $\mathcal{T}|A$ must be regrafted to either e_1 or e_2 . This completes the proof as in both cases θ is an NNI operation. \square

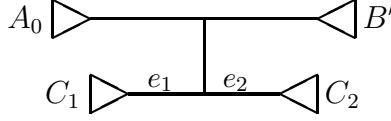


FIGURE 5. The tree \mathcal{T} for the case $k = 1$ in the proof of Lemma 3.1.

As a consequence of Lemma 3.1, we can express the sizes of both the SPR and the TBR neighbourhoods in terms of the number of each operation for a tree and the size of the NNI neighbourhood.

Lemma 3.2. *For $\mathcal{T} \in \mathcal{T}_n$, where $n \geq 4$, we have*

$$|N_{\text{SPR}}(\mathcal{T})| = |\mathcal{O}_{\text{SPR}}(\mathcal{T})| - 3|N_{\text{NNI}}(\mathcal{T})|,$$

and

$$|N_{\text{TBR}}(\mathcal{T})| = |\mathcal{O}_{\text{TBR}}(\mathcal{T})| - 3|N_{\text{NNI}}(\mathcal{T})|.$$

Proof. This follows from Lemma 3.1 and the observation in Section 2 that, if θ is an NNI operation for \mathcal{T} , then there are precisely four distinct operations $\theta' \in \mathcal{O}_{\text{NNI}}(\mathcal{T})$ such that $\theta(\mathcal{T}) = \theta'(\mathcal{T})$. \square

This lemma forms the basis of the two key results for this section. Both the number of distinct SPR operations and the number of distinct TBR operations for any given tree can be found relatively easily. We proceed with the SPR case first.

Theorem 3.3. *For a tree $\mathcal{T} \in \mathcal{T}_n$ where $n \geq 4$, we have*

$$|\mathcal{O}_{\text{SPR}}(\mathcal{T})| = 4(n-2)(n-3).$$

Proof. We consider two possible SPR operations on \mathcal{T} , firstly those that induce a trivial split on \mathcal{T} , and secondly those that induce a non-trivial split. In the first case, there are n possible leaves that can be pruned from \mathcal{T} , and for each leaf x there are $2n-6$ edges in $\mathcal{T}-x$ to which we can reconnect it so that the resulting tree is different from \mathcal{T} .

In the second case, suppose that the non-trivial split is $A|B$, with $|A| = a$ and $|B| = b$. If we choose $\mathcal{T}|A$ to be the pruned subtree, then there are $2b-3$ edges to which we can regraft $\mathcal{T}|A$. However, one of these results in the same tree as we began with, namely \mathcal{T} . Thus there are $2b-4$ such distinct operations. Similarly, if we choose $\mathcal{T}|B$ as the pruned subtree, then there are $2a-4$ possible SPR operations. Thus there are $2n-8$ distinct SPR operations for each of the $n-3$ non-trivial

splits of \mathcal{T} . Hence

$$\begin{aligned} |\mathcal{O}_{\text{SPR}}(\mathcal{T})| &= n(2n - 6) + (n - 3)(2n - 8) \\ &= 4(n - 2)(n - 3). \end{aligned}$$

□

As a corollary to this theorem, we obtain the result of Allen and Steel's [1] for the size of the SPR neighbourhood. The proof is omitted, as it follows trivially from the size of NNI neighbourhood (see Section 1), Lemma 3.2 and Theorem 3.3.

Corollary 3.4 (Theorem 2.1, [1]). *For $\mathcal{T} \in \mathcal{T}_n$ where $n \geq 4$, we have*

$$|N_{\text{SPR}}(\mathcal{T})| = 2(n - 3)(2n - 7).$$

We require one further idea before tackling the TBR problem. For a binary tree \mathcal{T} , we define $\Gamma(\mathcal{T})$ by

$$\Gamma(\mathcal{T}) = \sum |A| \cdot |B|,$$

where the sum is taken over all non-trivial splits $A|B$ of \mathcal{T} . This quantity is closely related to the Wiener index which arose out of chemical graph theory [2].

Theorem 3.5. *For a tree $\mathcal{T} \in \mathcal{T}_n$ where $n \geq 4$, we have*

$$|\mathcal{O}_{\text{TBR}}(\mathcal{T})| = 4\Gamma(\mathcal{T}) - 4(n - 2)(n - 3).$$

Proof. We consider two possible TBR operations on \mathcal{T} , firstly those that induce a trivial split on \mathcal{T} , and secondly those that induce a non-trivial split. The argument in the first case is identical to that given in the proof of Theorem 3.3, and gives $n(2n - 6)$ distinct TBR operations.

Now, let $A|B$ be some non-trivial split of \mathcal{T} induced by the edge e . Then when we bisect \mathcal{T} by deleting e , there are $2|A| - 3$ edges in one component of the resulting forest and $2|B| - 3$ edges in the other. Hence, there are $(2|A| - 3)(2|B| - 3)$ ways to choose an edge from each of $\mathcal{T}|A$ and $\mathcal{T}|B$. Precisely one of these results in re-forming \mathcal{T} . Hence, by taking a sum over all non-trivial splits $A|B$ of \mathcal{T} , we get

$$\begin{aligned} |\mathcal{O}_{\text{TBR}}(\mathcal{T})| &= n(2n - 6) + \sum [(2|A| - 3)(2|B| - 3) - 1] \\ &= 4\Gamma(\mathcal{T}) - 4(n - 2)(n - 3). \end{aligned}$$

□

This brings us to the following key result in this paper, which relates the size of the TBR neighbourhood of a phylogenetic tree to its shape, and provides an effective way to calculate this quantity. Also, as we will see in the next section, Theorem 3.6 gives us enough traction to

characterise the trees that respectively maximise and minimise the size of the TBR neighbourhood.

Theorem 3.6. *For $\mathcal{T} \in \mathcal{T}_n$ where $n \geq 4$, we have*

$$|N_{\text{TBR}}(\mathcal{T})| = 4\Gamma(\mathcal{T}) - (4n - 2)(n - 3).$$

Proof. This follows immediately from the size of the NNI neighbourhood, Lemma 3.2 and Theorem 3.5. \square

4. CHARACTERISATIONS OF THE EXTREMAL CASES

Since the size of the TBR neighbourhood for \mathcal{T} is dependent on both the number of leaves in \mathcal{T} and the shape of \mathcal{T} , it makes sense to characterise which tree shapes give the extreme values for this size. As a consequence of Theorem 3.6, it suffices to determine which tree shapes maximise and minimise the size of $\Gamma(\mathcal{T})$ over all trees in \mathcal{T}_n for some n . We begin with the easier case, that is, finding the trees that maximise $\Gamma(\mathcal{T})$.

Lemma 4.1. *Let $\mathcal{T} \in \mathcal{T}_n$ be a tree such that $\Gamma(\mathcal{T}) \geq \Gamma(\mathcal{T}')$ for all $\mathcal{T}' \in \mathcal{T}_n$. Then \mathcal{T} is a caterpillar.*

Proof. Suppose that $\{x_1, x_2\}$ and $\{x_3, x_4\}$ are cherries of \mathcal{T} , and let the sets Y_1, \dots, Y_k partition the remaining leaves so that \mathcal{T} can be represented as in Fig. 6.



FIGURE 6. The tree \mathcal{T} in the proof of Lemma 4.1.

Setting $y_i = |Y_i|$, it will suffice to show that $y_i = 1$ for all i . Assuming otherwise, let $i \in \{1, \dots, k\}$ be the smallest index such that $y_i > 1$. Now we form a second tree \mathcal{T}' by moving the subtree $\mathcal{T}|Y_i$ to the position adjacent to x_1 . The tree \mathcal{T}' is shown in Fig. 7.

Now, calculating the difference between $\Gamma(\mathcal{T})$ and $\Gamma(\mathcal{T}')$, we find that

$$\begin{aligned}
\Gamma(\mathcal{T}) - \Gamma(\mathcal{T}') &= \sum_{j=0}^{i-1} (j+2)(n-j-2) - \sum_{j=0}^{i-1} (y_i + j + 1)(n - y_i - j - 1) \\
&= \sum_{j=0}^{i-1} [(j+2)n - (j+2)^2 - (y_i + j + 1)n + (y_i + j + 1)^2] \\
&= (1 - y_i) \sum_{j=0}^{i-1} (n - y_i - 3 - 2j) \\
&= i(1 - y_i)(n - y_i - i - 2).
\end{aligned}$$

Since $y_j \geq 1$ for all j , we have the inequality $y_i + (i - 1) \leq n - 4$, from which $n - y_i - i - 2$ is strictly positive. Together with the assumption $y_i > 1$, we conclude that $\Gamma(\mathcal{T}) < \Gamma(\mathcal{T}')$, a contradiction as required. Therefore $y_i = 1$ indeed holds for all $i \in \{1, \dots, k\}$, and \mathcal{T} is a caterpillar.

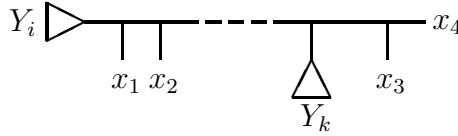


FIGURE 7. The tree \mathcal{T}' in the proof of Lemma 4.1.

□

Recall that the exact upper bound on the size of $N_{\text{TBR}}(\mathcal{T})$ for a tree $\mathcal{T} \in \mathcal{T}_n$ was proven in [3] by induction on n . Corollary 4.2 confirms this result using a different approach.

Corollary 4.2 (Theorem 2.1, [3]). *The tree $\mathcal{T} \in \mathcal{T}_n$ maximises the size of the TBR neighbourhood over \mathcal{T}_n if and only if \mathcal{T} is a caterpillar. Moreover, if \mathcal{T} is a caterpillar then*

$$|N_{\text{TBR}}(\mathcal{T})| = \frac{2}{3}n^3 - 4n^2 + \frac{16}{3}n + 2.$$

Proof. The first part of the corollary follows from Lemma 4.1. To find the size of the neighbourhood, we apply Theorem 3.6 from which we

have

$$\begin{aligned}
|N_{\text{TBR}}(\mathcal{T})| &= 4\Gamma(\mathcal{T}) - (4n - 2)(n - 3) \\
&= 4 \sum_{i=2}^{n-2} i(n - i) - (4n - 2)(n - 3) \\
&= \frac{2}{3}n^3 - 4n^2 + \frac{16}{3}n + 2.
\end{aligned}$$

□

Characterising the trees that minimise the size of the TBR neighbourhood relies heavily on Lemma 4.3. Before proving this, we give an example of the simplest case of this lemma. Referring to Fig. 8, suppose that the sizes of the pendant subtrees labelled by X_1, \dots, X_4 are x_1, \dots, x_4 respectively. If this tree has a minimal value for $\Gamma(\mathcal{T})$, then since $\Gamma(\mathcal{T})$ is the sum of $|A| \cdot |B|$ over all non-trivial splits $A|B$, we must have

$$(x_1 + x_2)(x_3 + x_4) \leq \min\{(x_1 + x_3)(x_2 + x_4), (x_1 + x_4)(x_2 + x_3)\}.$$

Assuming without loss of generality that x_1 is the smallest of the four quantities, it is easy to show that x_2 is the next smallest. Lemma 4.3 extends this observation to a more general result.

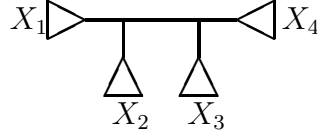


FIGURE 8. A tree illustrating the simplest case of Lemma 4.3.

Lemma 4.3. *Let $X = \{1, \dots, n\}$, and let $\mathcal{T} \in \mathcal{T}_n$ be such that $\Gamma(\mathcal{T}) \leq \Gamma(\mathcal{T}')$ for all $\mathcal{T}' \in \mathcal{T}_n$. Further, for some $k \geq 0$ let $X_1, \dots, X_4, Y_1, \dots, Y_k$ partition X such that the following hold:*

- (i) $X_i|(X - X_i) \in \Sigma(\mathcal{T})$ for all $i \in \{1, \dots, 4\}$;
- (ii) $Y_i|(X - Y_i) \in \Sigma(\mathcal{T})$ for all $i \in \{1, \dots, k\}$; and
- (iii) $A_i|(X - A_i) \in \Sigma(\mathcal{T})$ for all $i \in \{0, \dots, k\}$, where $A_0 = X_1 \cup X_2$, $A_i = A_{i-1} \cup Y_i$.

Then without loss of generality we have $x_1 \leq x_2 \leq x_3 \leq x_4$, where $x_i = |X_i|$.

Proof. Swapping the subscripts of X_i if necessary, we can assume $x_1 \leq x_2 \leq x_3$. Supposing that the lemma is false, we have $x_2 > x_4$. Then either $x_1 = x_3$, and so $x_1 \geq x_2 \geq x_3 \geq x_4$, contradicting our assumption that the lemma is false, or $x_1 < x_3$.

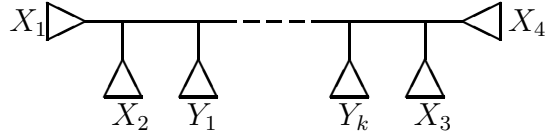


FIGURE 9. The tree \mathcal{T} in Lemma 4.3.

Figure 9 shows the general structure of a tree \mathcal{T} that satisfies the conditions of the lemma. Let \mathcal{T}_1 be the tree obtained from \mathcal{T} by swapping the subtrees labelled by X_1 and X_3 , and let \mathcal{T}_2 be similarly obtained by swapping the subtrees $\mathcal{T}|X_2$ and $\mathcal{T}|X_4$. Let $y_i = |Y_i|$, and $b_0 = 0, b_i = b_{i-1} + y_i$. Then we have

$$\begin{aligned} \Gamma(\mathcal{T}) - \Gamma(\mathcal{T}_1) &= \sum_{j=0}^k (x_1 + x_2 + b_j)(n - x_1 - x_2 - b_j) \\ &\quad - \sum_{j=0}^k (x_2 + x_3 + b_j)(n - x_2 - x_3 - b_j) \\ &= (x_3 - x_1) \left[2 \sum_{j=0}^k b_j - (k+1)(n - x_1 - 2x_2 - x_3) \right]. \end{aligned}$$

Since we assume that $\Gamma(\mathcal{T}) \leq \Gamma(\mathcal{T}_1)$, we get

$$\begin{aligned} \Gamma(\mathcal{T}) - \Gamma(\mathcal{T}_2) &= (x_4 - x_2) \left[2 \sum_{j=0}^k b_j - (k+1)(n - 2x_1 - x_2 - x_4) \right] \\ &> (x_4 - x_2) \left[2 \sum_{j=0}^k b_j - (k+1)(n - x_1 - 2x_2 - x_3) \right] \\ &= \frac{x_4 - x_2}{x_3 - x_1} (\Gamma(\mathcal{T}) - \Gamma(\mathcal{T}_1)) \\ &\geq 0, \end{aligned}$$

contradicting the fact that $\Gamma(\mathcal{T}) \leq \Gamma(\mathcal{T}_2)$. □

Applying Lemma 4.3, we can completely characterise those trees \mathcal{T} that minimise the size of $\Gamma(\mathcal{T})$, and therefore those trees that minimise the size of the TBR neighbourhood.

Lemma 4.4. *Let $X = \{1, 2, \dots, n\}$ for some $n = \sum_{i=0}^k \alpha_i 2^i$, where $\alpha_i \in \{0, 1\}$ for $0 \leq i < k$ and $\alpha_k = 1$. Let $\beta_j = \frac{1}{2^j} \sum_{i=j}^k \alpha_i 2^i$. Let $\mathcal{T} \in \mathcal{T}_n$ such that $\Gamma(\mathcal{T}) \leq \Gamma(\mathcal{T}')$ for all $\mathcal{T}' \in \mathcal{T}_n$. Then for all $0 \leq j \leq k-1$ there is a partition X_1, \dots, X_{β_j} of X into β_j disjoint subsets such that following properties hold:*

- (i) $X_p | (X - X_p) \in \Sigma(\mathcal{T})$ for all $1 \leq p \leq \beta_j$; and
- (ii) $|X_p| = 2^j$ for all $1 \leq p < \beta_j$.

Proof. For $j = 0$, this holds trivially. We assume that for some $0 \leq j < k-1$, the partition X_1, \dots, X_{β_j} of X satisfies the conditions of the lemma.

Suppose that for $1 \leq p < q < \beta_j$, there is no set Y that contains either X_p or X_q such that $Y | (X - Y) \in \Sigma(\mathcal{T})$ and $|Y| = 2^{j+1}$. Then we can apply Lemma 4.3 to find a tree \mathcal{T}' for which $\Gamma(\mathcal{T}') < \Gamma(\mathcal{T})$. Hence, for m such that $2m < \beta_j$, there are disjoint subsets X'_1, \dots, X'_m of X such that $X'_p | (X - X'_p) \in \Sigma(\mathcal{T})$ and $|X'_p| = 2^{j+1}$.

There are two cases to consider. Suppose firstly that $2m = \beta_j - 2$. Then there is some $1 \leq p < \beta_j$ such that X_p is not contained in some Y , where $Y | (X - Y) \in \Sigma(\mathcal{T})$ and $|Y| = 2^{j+1}$. We can then use Lemma 4.3 again to show that if $X'_{\beta_{j+1}} = X_p \cup X_{\beta_j}$, then $X'_{\beta_{j+1}} | (X - X'_{\beta_{j+1}}) \in \Sigma(\mathcal{T})$. Since $m + 1 = \beta_{j+1}$, we have the required partition.

On the other hand, if $2m = \beta_j - 1$ then we can use Lemma 4.3 to show that there is some $1 \leq p \leq m$ such that, if $X'_{\beta_{j+1}} = X_{\beta_j} \cup X'_p$, then $X'_{\beta_{j+1}} | (X - X'_{\beta_{j+1}}) \in \Sigma(\mathcal{T})$. Again, this gives the required partition, completing the induction. \square

The question now is what these trees look like. In some sense, the trees that minimise the size of $\Gamma(\mathcal{T})$ are maximally balanced, although we must carefully define what we mean by this. The only sizes of n for which an unrooted binary tree can be truly balanced, or perfect, are $n = 2^k$ or $n = 3 \cdot 2^k$, where the tree is vertex-transitive with respect to the leaves and we have either two-fold symmetry about an interior edge of the tree or three-fold symmetry about an interior vertex. For values of n other than those which admit a perfect tree, we necessarily lose the global property of leaf-transitivity.

A tree $\mathcal{T} \in \mathcal{T}_n$, where $3 \cdot 2^k \leq n < 3 \cdot 2^{k+1}$ for some $k \geq 0$, is called complete if and only if

- (i) there is a cluster Y of \mathcal{T} with $|Y| = 2^{k+1}$; and

- (ii) for all clusters Y with $2 \leq |Y| \leq 2^{k+1}$, there is a bipartition Y_1, Y_2 of Y such that both of Y_1, Y_2 are clusters of \mathcal{T} , and such that $|Y_1| = 2^j$ and $2^{j-1} \leq |Y_2| < 2^{j+1}$ for some j .

Intuitively, for each cluster Y with $|Y|$ being a power of 2, the pendant subtree $\mathcal{T}|Y$ is perfectly balanced. For more details on complete trees and a generalization of completeness to trees with arbitrary vertex degrees, see [4]. The trees in Lemma 4.4 are precisely the complete trees in the space \mathcal{T}_n , from which we obtain the next theorem. The proof is routine and omitted.

Theorem 4.5. *The tree $\mathcal{T} \in \mathcal{T}_n$ minimises the size of the TBR neighbourhood over \mathcal{T}_n if and only if \mathcal{T} is complete.*

Let us continue towards finding the size of the TBR neighbourhood for complete trees. To this end, we introduce one additional notation. For each positive integer m , there exists a unique binary expansion $m = \sum_{i=0}^k \alpha'_i 2^i$, where $\alpha'_i \in \{0, 1\}$ for $0 \leq i < k$ and $\alpha'_k = 1$. Let $\tau(m) = 1$ if $\alpha'_{k-1} = 1$, and $\tau(m) = 0$ otherwise. In particular, we have $\tau(2^k) = 0$ for every k .

Lemma 4.6. *Let $\mathcal{T} \in \mathcal{T}_n$ be a complete tree for some $n = \sum_{i=0}^k \alpha_i 2^i$, where $\alpha_i \in \{0, 1\}$ for $0 \leq i < k$ and $\alpha_k = 1$. Then:*

$$\begin{aligned} \Gamma(\mathcal{T}) = \sum_{j=1}^{k-1} \left[\left(\sum_{i=j}^k \alpha_i 2^i - 2^j \right) \left(2n - \sum_{i=j}^k \alpha_i 2^i \right) + \alpha_{j-1} 2^j (n - 2^j) \right] \\ + (\alpha_{k-1} - 1) 2^{k-1} (n - 2^{k-1}). \end{aligned}$$

Proof. We use the proof of Lemma 4.4 to obtain this result. For each of the partitions X_1, \dots, X_{β_j} , we take the sum of $|X_p| \cdot (n - |X_p|)$. Note that X_{β_j} contains a cluster of size 2^j if and only if $\tau(|X_{\beta_j}|) = 1$.

Consider a tree on n leaves where $\alpha_{k-1} = 1$ following the notation of Lemma 4.4. This gives

$$\begin{aligned} \Gamma(\mathcal{T}) &= \sum_{j=1}^{k-1} \left[\sum_{p=1}^{\beta_j} |X_p| \cdot (n - |X_p|) \right] + \tau(|X_{\beta_j}|) 2^j (n - 2^j) \\ &= \sum_{j=1}^{k-1} \left[2^j (\beta_j - 1 + \tau(|X_{\beta_j}|)) (n - 2^j) + |X_{\beta_j}| \cdot (n - |X_{\beta_j}|) \right]. \end{aligned}$$

We also have from Lemma 4.4 that

$$|X_{\beta_j}| = n - 2^j (\beta_j - 1),$$

and hence $\tau(|X_{\beta_j}|) = 1$ if and only if $\alpha_{j-1} = 1$. Incorporating this into the above expression, we find

$$\begin{aligned}\Gamma(\mathcal{T}) &= \sum_{j=1}^{k-1} [2^j(\beta_j - 1)(2n - 2^j\beta_j) + \alpha_{j-1}2^j(n - 2^j)] \\ &= \sum_{j=1}^{k-1} \left[\left(\sum_{i=j}^k \alpha_i 2^i - 2^j \right) \left(2n - \sum_{i=j}^k \alpha_i 2^i \right) + \alpha_{j-1} 2^j (n - 2^j) \right].\end{aligned}$$

In the case that $\alpha_{k-1} = 0$, the partition $X_1, \dots, X_{\beta_{k-1}}$ is a bipartition of the leaf set of \mathcal{T} , and so we need only take the product $|X_1| \cdot (n - |X_1|)$ once in the sum above. In other words, we need to subtract $2^{k-1}(n - 2^{k-1})$ from the formula. \square

We conclude this section with two corollaries, the first of which gives an exact value for the size of the TBR neighbourhood for perfect trees, and the second an asymptotic lower bound on the size of this neighbourhood for complete trees. Both proofs follow from Lemma 4.6 and Theorem 3.6.

Corollary 4.7. *Let $\mathcal{T} \in \mathcal{T}_n$ be a perfect tree. Then*

$$|N_{\text{TBR}}(\mathcal{T})| = n^2 \left(4k - \frac{32}{3} \right) + 22n - 6$$

if $n = 3 \cdot 2^{k-1}$ for some k , and

$$|N_{\text{TBR}}(\mathcal{T})| = n^2(4k - 13) + 22n - 6$$

if $n = 2^k$ for some k .

Proof. In the first case, where $n = 3 \cdot 2^{k-1}$, we have

$$\begin{aligned}\Gamma(\mathcal{T}) &= \sum_{j=1}^{k-1} n(n - 2^j) \\ &= n^2(k - 1) - n(2^k - 2) \\ &= n^2 \left(k - \frac{5}{3} \right) + 2n,\end{aligned}$$

and the result follows by applying Theorem 3.6. On the other hand, if $n = 2^k$ then

$$\begin{aligned}\Gamma(\mathcal{T}) &= \sum_{j=1}^{k-2} n(n - 2^j) + \frac{n^2}{4} \\ &= n^2 \left(k - \frac{7}{4} \right) - n(2^{k-1} - 2) \\ &= n^2 \left(k - \frac{9}{4} \right) + 2n,\end{aligned}$$

and again applying Theorem 3.6 gives the required result. \square

Corollary 4.8. *Let $\mathcal{T} \in \mathcal{T}_n$ be a complete tree. Then*

$$|N_{\text{TBR}}(\mathcal{T})| = 4n^2 \lfloor \log_2 n \rfloor + O(n^2).$$

Proof. The proof is similar in nature to that for the previous corollary. If $3 \cdot 2^{k-1} \leq n < 2^{k+1}$ for some $k \geq 1$, then we have

$$\begin{aligned}\Gamma(\mathcal{T}) &= \sum_{j=1}^{k-1} \left[\left(n - \sum_{i=0}^{j-1} \alpha_i 2^i - 2^j \right) \left(n + \sum_{i=0}^{j-1} \alpha_i 2^i \right) + \alpha_{j-1} 2^j (n - 2^j) \right] \\ &= n^2(k-1) - n(2^k - 2) + \sum_{j=1}^{k-1} \alpha_{j-1} 2^j (n - 2^j) - \sum_{j=1}^{k-1} \left(\sum_{i=0}^{j-1} \alpha_i 2^i \right)^2.\end{aligned}$$

However, we can obtain a bound for the final term of this expression by assuming that $\alpha_i = 1$ for all $i \in \{0, \dots, k-2\}$, giving

$$\begin{aligned}\sum_{j=1}^{k-1} \left(\sum_{i=0}^{j-1} \alpha_i 2^i \right)^2 &< \sum_{j=1}^{k-1} 2^{2j} \\ &= \frac{2}{3} (2^{2k-1} - 1) \\ &= O(n^2).\end{aligned}$$

Similarly, we have $\sum_{j=1}^{k-1} \alpha_{j-1} 2^j (n - 2^j) = O(n^2)$, as required.

The other case, where $2^k \leq n < 3 \cdot 2^{k-1}$, follows in a similar manner, and we complete the proof by applying Theorem 3.6. \square

REFERENCES

- [1] Allen, B.L. and Steel, M. (2001). Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, **5**, 1–15.
- [2] Dobrynin, A.A., Entringer, R. and Gutman, I. (2001). Wiener index of trees: theory and applications. *Acta Applicandae Mathematicae*, **66**, 211–249.

- [3] Humphries, P.J. (2011). Bounds on the size of the TBR unit-neighbourhood. *Annals of Combinatorics*, **14**, 479–485.
- [4] Szekely, L., Wang, H. and Wu T. (2011). The sum of the distances between the leaves of a tree and the semi-regular property. *Discrete Mathematics*, **311**, 1197–1203.
- [5] Katoh, K., and Miyata, T. (1999). A heuristic approach of maximum likelihood method for inferring phylogenetic tree and an application to the mammalian SOX-3 origin of the testis-determining gene SRY. *FEBS Letters*, **463**, 129–132.
- [6] Kubatko, L.S. (2008). Inference of Phylogenetic Trees, in *Tutorials in Mathematical Biosciences IV: Evolution and Ecology*, A. Friedman (ed.), Springer-Verlag, 1–38.
- [7] Robinson, D.F. (1971). Comparison of Labeled Trees with Valency Three. *Journal of Combinatorial Theory*, **11**, 105–119.

DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF CANTERBURY, CHRISTCHURCH, NEW ZEALAND

E-mail address: pjhumphries@gmail.com

SCHOOL OF COMPUTING SCIENCES, UNIVERSITY OF EAST ANGLIA, NORWICH, UNITED KINGDOM

E-mail address: taoyang.wu@gmail.com